INSTITUTE FOR DEVELOPMENT STUDIES
UNIVERSITY OF NAIROBI

Discussion Paper No. 100

LEAST SQUARES ESTIMATION OF RELATIONS AND SYSTEMS OF RELATIONS INVOLVING CATEGORICAL DEPENDENT VARIABLES

by

Gary S. Fields

November, 1970

LEAST SQUARES ESTIMATION OF RELATIONS AND SYSTEMS OF RELATIONS INVOLVING CATEGORICAL DEPENDENT VARIABLES

Gary S. Fields

November, 1970

I. INTRODUCTION

The purpose of this paper is to consider statistical estimation of relations in which the dependent variable(s) is (are) categorical, taking on one of a number of discrete values. The results herein constitute a generalization of the work of Zellner and others in cases where the dependent variable or variables may have only two possible values.

Categorical dependent variables may occur naturally or may be created artificially in situations in which only a few values are possible. An example of a natural occurrence would be where the dependent variable is family ownership of a particular durable, say refrigerators, where the only possibilities with any frequency are 0, 1, or 2. An artificial situation involves the scaling of a phenomenon, such as voter attitudes pro, neutral, or con toward a public project. The arbitrary numbers +1, 0, and -1 might be chosen. Both these cases are examples of "trichotomous" variables. Categorical variables with two possibilities are said to be "dichotomous", those with many possibilities "multichotomous."

The plan of this paper is to consider in Section II the simplest case, that of a single equation model with a dichotomous dependent variable. These results are extended in Section III to include cases involving a multichotomous dependent variable. In Section IV, Zellner and Lee's full information estimation procedure for a system of categorical relations is considered and applied to the case of a model with several multichotomous dependent variables. Empirical examples using data on retraining the hard-core unemployed in the United States are presented in Section V. The results are summarized in Section VI.

II. SINGLE EQUATION ESTIMATION WHEN THE DEPENDENT VARIABLE IS DICHOTOMOUS

Suppose we hypothesize a linear model in which the dependent variable is dichotomous. Plausible circumstances for such a relation involve the occurrence or non-occurrence of a particular event, membership in one or the other of two mutually exclusive groups, and similar situations where the dependent variable can be characterized by an either/or classification. Although the dependent variable could be assigned any two arbitrary values, the values one and zero permit the interpretation of the regression model as a "linear probability function." This terminology is used by Warner [1], Ladd [2], and others.

Considering the dependent variable as membership in one of two classes, the original model states that the probability of membership in the first class is a linear combination of the regressors, plus an error term:

$$(1) \quad \underset{(N\times 1)}{y} = \underset{(N\times M)}{X} \underset{(M\times 1)}{\beta} + \underset{(N\times 1)}{\varepsilon}.$$

If it is assumed that

$$(2)\ E(\varepsilon) = 0,$$

$$(3)\ \mathrm{Var}_{\varepsilon} = E(\varepsilon\varepsilon') = \sigma^2 I,$$

and (4) X and ε are independent,

then the ordinary least squares estimator,

$$(5)\ b_{OLS} = (X'X)^{-1}X'y$$

is unbiased. The predicted value of the dependent variable

---

[1] Stanley Warner, Stochastic Choice of Mode in Urban Travel: A Study in Binary Choice, Northwestern University Press, 1962

[2] George W. Ladd, "Linear Probability Functions and Discriminant Functions," Econometrica, October, 1966

conditional on a particular value of the X's, $\hat{y}|$ Xo, is an unbiased estimate of the conditional probability of membership in the class in question.

A serious difficulty with the ordinary least squares formulation is that the dependent variable y, the probability of membership in the first class, is not measureable. Instead, for each individual, we have available a figure 1 or 0, which represents the ex post probability. Our problem is then of the errors in observation type. Let Y represent the vector of observed (ex post) probabilities. The true values and observed values are related by

$$(6) \quad Y = y + \theta$$

where θ is a vector of measurement errors. Under the assumption that the X's are correctly measured, (6) can be substituted into (1), yielding

$$(7) \quad Y = X\beta + (\varepsilon + \theta).$$

Note that the error term is a composite one, including a true stochastic component and a measurement error. Goldberger [3] fails to distinguish this composite term from the original error $\in$.

Let us assume that

$$(8) \quad E\theta = 0.$$

By (2) and (8), the expected value of the composite error term in (7) equals zero, or $(9) \quad E (\varepsilon + \theta) = 0.$

Focusing on the i'th individual, the composite error term $(\varepsilon_i + \theta_i)$ has only two possible values, the frequencies being determined to conform with (9):

$$\begin{array}{cc} \dfrac{(\varepsilon_i + \theta_i)}{1 - X_i\beta} & \dfrac{f(\varepsilon_i + \theta_i)}{X_i\beta} \\ -X_i\beta & 1 - X_i\beta \end{array}$$

The variance of the composite error term is

$$E\{(\varepsilon_i + \theta_i)(\varepsilon_i + \theta_i)'\} = (1-X_i\beta)^2 X_i\beta + (-X_i\beta)^2(1-X_i\beta)$$
$$= X_i\beta(1-X_i\beta)\{(1-X_i\beta) + X_i\beta\} = X_i\beta(1-X_i\beta) = EY_i(1-EY_i).$$

The variance is seen to be a function of EY_i and therefore of X_i. Thus, although the original model assumed homoscedasticity in (3), that assumption cannot be maintained in an empirical testing situation in which measurement

---

[3] Arthur S. Goldberger, Econometric Theory, Wiley, 1964, pp.248-250

errors are inevitable.

Having established the existence of heteroscedasticity, what would happen if we then proceeded to ignore it? The ordinary least squares estimate of β in (7) is $b_{OLS} = (X'X)^{-1}X'Y$. This estimate is still unbiased.[4] However, b_{OLS} is no longer the minimum variance estimate.[5] In addition, the classical estimator of σ^2, $s^2 = E(e'e)/N-K$, is biased.[6] The classical estimate of the covariance matrix of b_{OLS} is $E(b_{OLS} - \beta)(b_{OLS} - \beta)' = s^2(X'X)^{-1}$, which is biased since $Es^2 \neq \sigma^2$. Failure to consider heteroscedasticity would thus yield a b_{OLS} which is an unbiased estimator of β but result in a biased estimate of the variance of b_{OLS}.

The heteroscedasticity resulting from errors in measurement may be taken into consideration by constructing a new model

(7) $Y = X\beta + (\varepsilon + \theta)$

under assumptions

(9) $E(\varepsilon + \theta) = 0$

(10) $Var(\varepsilon + \theta) = \sigma^2 D$, D diagonal, $D \neq I$,

(4) X and ε are independent

(11) X and θ are independent.

For empirical testing the heteroscedasticity may be accounted for in the standard manner by

1. approximating EY_i by $\hat{Y}_i$, the ordinary least squares estimate,
2. estimating the variance/covariance matrix by

(12) $$\hat{D} = \begin{bmatrix} \hat{Y}_1(1-\hat{Y}_1) & 0 & \cdots & 0 \\ 0 & \hat{Y}_2(1-\hat{Y}_2) & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \hat{Y}_n(1-\hat{Y}_n) \end{bmatrix}$$

4 $E(b_{OLS}) = E\left[(X'X)^{-1}X'(X\beta+\varepsilon)\right] = \beta + E\{(X'X)^{-1}X'\varepsilon\} = \beta$, since X and ε are independent by (4).

5 In the general linear model, where the variance-covariance matrix is $\sigma^2 D$, $D \neq I$, it is a standard result that the quasi-generalized least squares estimator

$$b_{GLS} = (X'D^{-1}X)^{-1}X'D^{-1}Y$$

is the maximum likelihood estimator of β. Under very general regularity conditons, the maximum likelihood estimator is asymptotically efficient. Hence, any other unbiased estimator, in particular the ordinary least squares estimator b_{OLS}, is less efficient.

6 $E(e'e) = tr\{A\ E(\varepsilon+\theta\ )(\varepsilon+\theta\ )'\}$.
where $A = I - X(X'X)^{-1}X'$.
$E\{(\varepsilon+\theta\ )(\varepsilon+\theta)'\} = \sigma^2 D$.
Therefore, $E(e'e) = \sigma^2 tr\{D - X(X'X)^{-1}X'D\}$
$= \sigma^2 trD - \sigma^2 tr\ X(X'X)^{-1}X'D$
$= \sigma^2\{trD - tr((X'X)^{-1}X'DX)\}$.
Therefore, $Es^2 = E(e'e)/N-M \neq \sigma^2$ in general.

and 3. using this matrix to construct a quasi-generalized least squares (Aitken) estimator

$$(13)\quad b_{GLS} = (X'\hat{D}^{-1}X)^{-1}X'\hat{D}^{-1}Y.$$

Since $\hat{D}$ is only an approximation to the true D, b_{GLS} is not BLUE. However, it does make allowance for heteroscedasticity and is therefore superior to the OLS estimator.

All the required results may be obtained by running OLS on a transformed model. A basic theorem of linear models is that for any symmetric positive definite matrix D, there exists a non-singular matrix T such that $TDT' = I$ and $T'T = D^{-1}$. The T corresponding to the D in (10) is

$$(14)\quad T_i = (D_i)^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{EY_1(1-EY_1)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{EY_2(1-EY_2)}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{EY_n(1-EY_n)}} \end{bmatrix}$$

Our matrix $\hat{D}$ is diagonal, so it is clearly symmetric; each element is the product of two numbers between 0 and 1, so it is positive definite. There then exists a matrix $\hat{T}$ such that $\hat{T}_i = (\hat{D}_i)^{-\frac{1}{2}}$. $\hat{T}$ is the (NxN) matrix

$$\hat{T} = \begin{bmatrix} \frac{1}{\sqrt{\hat{Y}_1(1-\hat{Y}_1)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\hat{Y}_2(1-\hat{Y}_2)}} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\hat{Y}_n(1-\hat{Y}_n)}} \end{bmatrix}$$

Clearly, $\hat{T}\hat{D}\hat{T}' = I$ and $\hat{T}\hat{T}' = \hat{D}^{-1}$.

If the linear model (7) is premultiplied by $\hat{T}$, we obtain:

$$(15)\quad \hat{T}Y = \hat{T}X\beta + \hat{T}(\varepsilon + \theta).$$

$$E\{\hat{T}(\varepsilon + \theta)\} = \hat{T}E(\varepsilon + \theta) = 0$$

$$\text{and Cov}\{\hat{T}(\varepsilon + \theta)\} = E\{\hat{T}(\varepsilon + \theta)((\hat{T}(\varepsilon + \theta)'\} = E\{\hat{T}(\varepsilon + \theta)(\varepsilon + \theta)'\hat{T}'\}$$

$$= \hat{T}\,E\{(\varepsilon + \theta)(\varepsilon + \theta)'\}\hat{T}' = \sigma^2\hat{T}\hat{D}\hat{T}'$$

$$= \sigma^2 I.$$

(15) therefore satisfies the Gauss-Markov assumptions, so ordinary least squares can be applied to the transformed data with optimal results. The resultant set of OLS regression coefficients may be shown equivalent to the Aitken estimator (13). The OLS estimator on the transformed model is $b_* = \{(\hat{T}X)'\hat{T}X\}^{-1}(\hat{T}X)'\hat{T}Y$, or

$$(16)\quad b_* = (X'\hat{T}'\hat{T}X)^{-1}X'\hat{T}'\hat{T}Y.$$

$$\hat{T}'\hat{T} = \begin{bmatrix} \frac{1}{Y_1(1 - Y_1)} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{Y_N(1 - Y_N)} \end{bmatrix} = \hat{D}^{-1}.$$

Substituting $\hat{D}^{-1}$ for $\hat{T}'\hat{T}$ in (16), we obtain

$$(17)\quad b_* = (X'\hat{D}^{-1}X)^{-1}(X'\hat{D}^{-1}Y),$$

which is identical with the Aitken estimator (13). The variance/covariance matrix of b_* is

$$(18)\quad E\{(b_* - \beta)(b_* - \beta)'\}$$

$$= E\{((\hat{T}X)'\hat{T}X)^{-1}(\hat{T}X)'\hat{T}(\varepsilon + \theta)(\varepsilon + \theta)'\hat{T}'\hat{T}X((\hat{T}X)'\hat{T}X)^{-1}\}$$

$$= E\{(X'\hat{T}'\hat{T}X)^{-1}X'\hat{T}'\hat{T}(\varepsilon + \theta)(\varepsilon + \theta)'\hat{T}'\hat{T}X(X'\hat{T}'\hat{T}X)^{-1}\}$$

$$= \sigma^2 E\{(X'\hat{T}'\hat{T}X)^{-1}(X'\hat{T}'\hat{T}X)(X'\hat{T}'\hat{T}X)^{-1}\}$$

$$= \sigma^2 E\{(X'\hat{T}'\hat{T}X)^{-1}\}$$

$$= \sigma^2 (X'\hat{D}^{-1}X)^{-1},$$

which is identical with the variance/covariance matrix using the Aitken estimator. The unbiased estimator of σ^2 from OLS on the transformed data is identical with that obtained by Aitken estimation. Define a vector of residuals on the transformed model: $\hat{T}e = \hat{T}Y - \hat{T}Xb_*$. The unbiased estimator of σ^2 from OLS is $\frac{(\hat{T}e)'\hat{T}e}{N - M}$, which is $\frac{e'\hat{T}'\hat{T}e}{N - M} = \frac{e'\hat{D}^{-1}e}{N - M}$, which is the unbiased estimator from the Aitken procedure. This unbiased estimator may be substituted into (18) to yield an estimator of the covariance matrix of b_*. Thus there is no need to calculate generalized least squares results (for which computer programs are generally unavilable), since all required estimates are obtainable in terms of OLS results.

III. EXTENSION TO THE CASE OF A MULTICHOTOMOUS DEPENDENT VARIABLE

The results derived in the previous section may readily be extended to the case of a trichotomous dependent variable in an analogous manner. Trichotomous variables are unusual in economics. One plausible situation arises in the use of published data when all but two categories are merged and assigned some arbitrary value. For instance, family automobile ownership might take on the values 0, 1, or 2.3, where 2.3 represents the mean automobile ownership of those who own at least two cars. Another plausible circumstance is where the dependent variable is of the pro, con, or neutral type. In that case an implicit assumption is made that the numerical difference between pro and neutral bears the same relation to the numerical difference between con and neutral as the relative "distances" in the world. For example, a 1, 2, 3 scale implicitly assumes that a pro response and a con response are equally different from neutrality.

Suppose we have the same model as in the previous case except that there are now three possible values of the dependent variable: Y_1, Y_2, and Y_3. Suppose further that the frequencies of these values for individuals with independent variable vector X_i are respectively f_1, f_2, and f_3. The composite error term for the i'th individual has three possible values with corresponding frequencies

$\varepsilon_i + \theta_i$	$f(\varepsilon_i + \theta_i)$
$Y_1 - X_i\beta$	f_1
$Y_2 - X_i\beta$	f_2
$Y_3 - X_i\beta$	f_3

By the assumption that the expected error is zero,

$$(19)\quad f_1(Y_1 - X_i\beta) + f_2(Y_2 - X_i\beta) + f_3(Y_3 - X_i\beta) = 0.$$

Furthermore,

$$(20)\quad f_1 + f_2 + f_3 = 1.$$

Combining (19) and (20) yields

$$(21)\quad f_1 Y_1 + f_2 Y_2 + f_3 Y_3 = X_i\beta$$

which says that the expected value of Y equals $X_i\beta$. The variance of the composite error term is

$$\begin{aligned} E(\varepsilon_i + \theta_i) &= f_1(Y_1 - X_i\beta)^2 + f_2(Y_2 - X_i\beta)^2 + f_3(Y_3 - X_i\beta)^2 \\ &= f_1\left[Y_1^2 - 2Y_1X_i\beta + (X_i\beta)^2\right] + f_2\left[Y_2^2 - 2Y_2X_i\beta + (X_i\beta)^2\right] \\ &\quad + f_3\left[Y_3^2 - 2Y_3X_i\beta + (X_i\beta)^2\right] \\ &= (f_1Y_1^2 + f_2Y_2^2 + f_3Y_3^2) - 2X_i\beta(f_1Y_1 + f_2Y_2 + f_3Y_3) \\ &\quad + (f_1 + f_2 + f_3)(X_i\beta)^2 \\ &= X_i\beta - 2X_i\beta(X_i\beta) + (X_i\beta)^2 = X_i\beta(1 - X_i\beta) \end{aligned}$$

or $$(22)\quad \mathrm{Var}(\varepsilon_i + \theta_i) = EY_i(1 - EY_i).$$

This variance is identical with the variance of the composite error term in the dichotomous case and varies systematically with X_i. The nature of the heteroscedasticity is therefore the same when the dependent variable is trichotomous as when it is dichotomous. However, the appropriate correction and estimation procedures are slightly different. For the matrix $\hat{T}$ to be non-imaginary, each number whose square root is to be taken must be positive. This is assured only when the predicted values are constrained to lie between zero and one. The dependent variable should be rescaled so that all predicted values fall within the required range. The matrix $\hat{T}$ will then exist, the computations can be performed, and the dependent variable can be returned to its original scale simply by multiplying each regression coefficient by the appropriate scaling factor.

The results presented in this section may easily be generalized to cases where the dependent variable has more than three possible values. The heteroscedasticity correction is identical.

It should be remarked that in cases where the Aitken generalized least squares estimation procedure is used to correct for heteroscedasticity due to a categorical dependent variable, the coefficient of determination R^2 will appear to increase substantially. This does not necessarily indicate a better fit. This is because the transformed model (15) has no intercept, and the regression line is therefore constrained to pass through the origin. The R^2 is calculated around the origin rather than the means, and will usually be higher.

IV. ESTIMATION IN MULTIVARIATE MULTIPLE REGRESSION WHEN THE DEPENDENT VARIABLES ARE MULTICHOTOMOUS ---THE ZELLNER-LEE FULL INFORMATION METHOD

Multivariate multiple regression [7] is a procedure for estimating the parameters of a system of equations in which each dependent variable is a function of one or more exogenous variables, none of which is the dependent variable in any other equation. In other words, the set of dependent variables in the several equations and the set of independent variables are mutually exclusive. Such equations may appear "seemingly unrelated," but if there is covariance in the errors between equations, a simultaneous equations approach is required for efficient estimation.

The discussion in this section focuses on a two-equation system in which each of the dependent variables is multichotomous. Extension to the case of more than two equations is straight-forward and is done at the end of the section. The particular stochastic assumptions in this discussion are those which are particularly relevant for cross-sectional survey data; consequently, the language used is appropriate for a study in which the dependent variable is the individual's response to survey questions.

Let there be observations on N cases (persons) where the dependent variables are Y_1 and Y_2. [8] Each dependent variable is assumed to enter into one and only one equation. Let X_1 and X_2 be $N \times M_e$ ($e = 1, 2$) matrices of observations on the independent variables in the e'th equation; let β_1 and β_2 be $M_e \times 1$ ($e = 1, 2$) vectors of regression coefficients, and ε_1 and ε_2 be $N \times 1$ vectors of errors. The system of equations may then be written

---

[7] Goldberger terms this "The Multivariate Contemporaneously Uncorrelated Linear Regression Model", and Zellner "Seemingly Unrelated Regressions."

[8] Note that Y_1 and Y_2 are now vectors corresponding to different equations, while in the previous sections they were scalars corresponding to different individuals. The same is true of the X's and ε's . Note also that for convenience, the composite error term $(\varepsilon_i + \theta_i)$ is being written simply as ε_i.

$$(23)\qquad \underset{(N\times 1)}{Y_1} = \underset{(N\times K_1)}{X_1}\ \underset{(K_1\times 1)}{\beta_1} + \underset{(N\times 1)}{\varepsilon_1}$$

and $$(24)\qquad \underset{(N\times 1)}{Y_2} = \underset{(N\times K_2)}{X_2}\ \underset{(K_2\times 1)}{\beta_2} + \underset{(N\times 1)}{\varepsilon_2}$$

The stochastic assumptions of the model are

$$(25)\qquad E\,\varepsilon_e = 0,\quad e = 1,2$$

$$(26)\qquad E\,\varepsilon_e\,\varepsilon_e' = D_{ee},\ \text{where } D_{ee} \text{ is an } N\times N \text{ diagonal matrix and } D_{ee} \neq I$$

and $$(27)\qquad E\,\varepsilon_e\,\varepsilon_f' = D_{ef},\ e \neq f,\ \text{where } D_{ef} \text{ is an } N\times N \text{ diagonal matrix and } D_{ef} \neq I.$$

(25) says that the vector of disturbances has expected value of zero. (26) says that the responses of different persons are independent but that the variances of responses on a given question for different persons are not equal (heteroscedasticity). (27) says that there is covariance between equations for a given individual, but no covariance between individuals; furthermore, the covariance is also heteroscedastic. This inter-equational dependency assumed in (27) necessitates a simultaneous equations approach for efficient estimation.[9]

Equations (23) and (24) may be written compactly as (28) $Y = XB + \varepsilon$

where $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ $X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}$ $B = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$.

Combining (26) and (27), the variance/covariance matrix is seen as

$$(29) \quad E\,\varepsilon\varepsilon' = \begin{bmatrix} E\varepsilon_1\varepsilon_1' & E\varepsilon_1\varepsilon_2' \\ E\varepsilon_2\varepsilon_1' & E\varepsilon_2\varepsilon_2' \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \Omega \ .$$

Following Zellner and Lee, the model may then be viewed as a multivariate multiple regression equation system (28) subject to stochastic specifications (25) and (29); statistical estimation may then be carried out by means of Aitken estimation.

Consider first the case of a dichotomous dependent variable where one and zero are the possible values. Let X_{i1} and ε_{i1} respectively be the vectors of independent variables and residuals in the first equation for the i'th individual, and β_1 the vector of coefficients on the independent variables. The residual on the first equation can take on two possible values with frequencies f_{11} and f_{21}. Adopting similar notation for the second equation,

---

[9] The gain in efficiency is discussed by Zellner in "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, June, 1962. In Zellner's words, "Since the Aitken estimator.... differs from that derived by application of least-squares equation-by-equation, it must be the case that the Aitken estimator is more efficient. Essentially, this gain in efficiency occurs because in estimating the coefficients of a single equation, the Aitken procedure takes account of zero restrictions on coefficients occurring in other equations." (p. 353)

the frequency distributions of the residuals in the two equations are:

ε_{i1}	$f(\varepsilon_{i1})$	ε_{i2}	$f(\varepsilon_{i2})$
$1 - X_{i1}\beta_1$	f_{11}	$1 - X_{i2}\beta_2$	f_{12}
$- X_{i1}\beta_1$	f_{21}	$- X_{i2}\beta_2$	f_{22}

The upper left block of Ω, D_{11}, is identical with the D matrix in the previous two sections, the i'th element of which is (30) $EY_{1i}(1 - EY_{1i})$. (See equation (22)). Similarly, the i'th element of the lower right block is $EY_{2i}(1 - EY_{2i})$. It remains now to solve for the i'th element of the off-diagonal term. Looking at the lower left and letting P_{jk} be the probability of being in the j'th category of the first equation and the k'th category of the second, the i'th element is

$$\begin{aligned}(31)\ E\varepsilon_{i2}\varepsilon_{i1} &= P_{11}(1 - X_{i1}\beta_1)(1 - X_{i2}\beta_2) + P_{12}(1-X_{i1}\beta_1)(-X_{i2}\beta_2) \\ &\quad + P_{21}(-X_{i1}\beta_1)(1 - X_{i2}\beta_2) + P_{22}(-X_{i1}\beta_1)(-X_{i2}\beta_2) \\ &= EY_{i1}Y_{i2} - EY_{i1}Y_{i2} - EY_{i1}EY_{i2} + EY_{i1}Y_{i2} \\ &= EY_{i1}Y_{i2} - EY_{i1}EY_{i2}\end{aligned}$$

Elements of the type (30) and (31) thus comprise the true covariance matrix

Zellner and Lee express their results in slightly different form by grouping cases with identical X vectors and by recognizing that in the 1/0 dichotomous case the proportions and expected values are identical,[10]

Consider now the case where at least one of the dependent variables is trichotomous. Let Y_{11}, Y_{21}, and Y_{31} be the possible values of the dependent variable in the first equation and Y_{12}, Y_{22}, and Y_{32} in the second. The frequency distributions in the two equations are

ε_{i1}	$f(\varepsilon_{i1})$	ε_{i2}	$f(\varepsilon_{i2})$
$Y_{11} - X_{i1}\beta_1$	f_{11}	$Y_{12} - X_{i2}\beta_2$	f_{12}
$Y_{21} \quad X_{i1}\beta_1$	f_{21}	$Y_{22} \quad X_{i2}\beta_2$	f_{22}
$Y_{31} \quad X_{i1}\beta_1$	f_{31}	$Y_{32} \quad X_{i2}\beta_2$	f_{32}

---

10 The reason that cases are not grouped in this paper is that in many, perhaps most, economic contexts, at least one independent variable is continuous. The only independent variable in Zellner and Lee's example is family income from categorical responses. Had income been available to the nearest hundred dollars or had they also included additional regressors, the number of groups would have quickly become unwieldy.

As shown in Section II, the elements of D_{11} are of the form of (30). The lower left element analogous to (31) is

$$(32)\quad E\varepsilon_{i2}\varepsilon_{i1} = \sum_{j=1}^{3}\sum_{k=1}^{3} P_{jk}(Y_{j1} - X_{i1}\beta_1)(Y_{k2} - X_{i2}\beta_2)$$

$$= \sum_{j=1}^{3}\sum_{k=1}^{3} P_{jk}(Y_{j1}Y_{k2} - X_{i1}\beta_1 Y_{k2} - Y_{j1}X_{i2}\beta_2 + X_{i1}\beta_1 X_{i2}\beta_2)$$

$$= \sum_{j=1}^{3}\sum_{k=1}^{3} P_{jk}(-Y_{j1}X_{i2}\beta_2) + \sum_{j=1}^{3}\sum_{k=1}^{3} P_{jk}(X_{i1}\beta_1 X_{i2}\beta_2)$$

$$+ \sum_{j=1}^{3}\sum_{k=1}^{3} P_{jk}(-Y_{j1}X_{i2}\beta_2) + \sum_{j=1}^{3}\sum_{k=1}^{3} P_{jk}(X_{i1}\beta_1 X_{i2}\beta_2)$$

$$= EY_{i1}Y_{i2} - EY_{i1}EY_{i2} - EY_{i1}EY_{i2} + EY_{i1}EY_{i2}$$

$$= EY_{i1}Y_{i2} - EY_{i1}EY_{i2}$$

which is exactly the same as (31). Elements of the type (30) and (32) comprise the true covariance matrix Ω in the trichotomous case.

Extension to larger systems of equations and dependent variables with more possibilities is straightforward. Suppose that there are E equations and each dependent variable has as many as π possible values. The equation system is

$$Y_1 = X_1\beta_1 + \varepsilon_1$$
$$Y_2 = X_2\beta_2 + \varepsilon_2$$
$$\vdots$$
$$Y_E = X_E\beta_E + \varepsilon_E$$

or $$(33)\quad Y = XB + \varepsilon$$

where $$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_E \end{bmatrix},\quad X = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & X_E \end{bmatrix},\quad B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_E \end{bmatrix},\quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_E \end{bmatrix}$$

The variance/covariance matrix is

$$(34)\quad E\varepsilon\varepsilon' = \begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1E} \\ D_{21} & D_{22} & \cdots & D_{2E} \\ \vdots & \vdots & D_{ce} & \vdots \\ D_{E1} & D_{E2} & \cdots & D_{EE} \end{bmatrix}$$

The frequency distribution of the stochastic term for the i'th individual on the e'th equation is

(35)

ε_{ie}	$f(\varepsilon_{ie})$	
$Y_{1e} - X_{1e}\beta_e$	f_{1e}	
$Y_{2e} - X_{2e}\beta_e$	f_{2e}	$e = 1, 2, \ldots E.$
$\vdots$	$\vdots$	
$Y_{\pi e} - X_{\pi e}\beta_e$	$f_{\pi e}$	

The i'th element of D_{ee} is

$$(36)\quad EY_{ie}(1 - EY_{ie})$$

and the i'th element of D_{ef}, $e \neq f$ is

$$(37)\quad E\varepsilon_{ie}\varepsilon_{if}$$

$$= \sum_{j=1}^{\pi}\sum_{k=1}^{\pi} P_{jk}(Y_{je}-X_{ie}\beta_e)(Y_{kf}-X_{if}\beta_f)$$

$$= \sum_{j=1}^{\pi}\sum_{k=1}^{\pi} P_{jk}(Y_{je}Y_{kf}-X_{ie}\beta_eY_{kf}-Y_{je}X_{if}\beta_f+X_{ie}\beta_eX_{if}\beta_f)$$

$$= EY_{ie}Y_{if} - EY_{ie}EY_{if}.$$

The Aitken generalized least squares estimator for the multivariate multiple regression system (33) is

$$(38)\quad b_{AIM} = (X'\Omega^{-1}X)^{-1}X\Omega^{-1}Y,$$

where Ω is the matrix of variances and covariances described in the previous paragraphs and X and Y are matrices as defined in (33). However, Ω is not ordinarily known. How can a consistent estimate of Ω be obtained? The procedure suggested by Zellner and Lee is to group the observations with the same X's, estimate the expected value of each Y_i by weighted least squares equation-by-equation, and estimate the intersection probabilities P_{jk} from the sample proportions in each group. However, this procedure breaks down either when one of the independent variables is continuous or when there are several categorical independent variables. In either case the number of groups approaches the sample size. Consequently the Zellner-Lee estimates of P_{jk} are based on only one or a few observations and can hardly be expected to approximate the elements of the true variance/covariance matrix (34). (See footnote 9.) The same holds for the cross-expect tions in the multichotomous case.

A procedure that will yield an unbiased estimator of Ω in such circumstances is the following. Consider the system of equations (33) under the maintained stochastic assumptions (25), (26), and (27). Assume also that X_e and ε_e are independent. Then an unbiased estimator of (36) is given by

$$(39) \qquad \hat{Y}_{ie}\,(1 - \hat{Y}_{ie}),$$

where the $\hat{Y}$'s are the predicted values from OLS on the e'th equation.[11] If Y_e and Y_f are independent random variables, then $EY_eY_f = EY_eEY_f$, so that zero is an unbiased estimate of (37). In a two-equation model, this is a case where multivariate multiple regression estimation gives no gain in efficiency over Aitken estimation equation-by-equation. However, in cases where Y_e and Y_f are dependent, the relationship can be described more or less accurately by a regression line. Suppose the relationship is approximately a linear one, so that $Y_f = \gamma + \delta Y_e + \eta$, and suppose further that Y_{ie} and η_i are independent and $E\eta_i = 0$ for all i. Let c and d be the OLS estimates of γ and δ. Then an unbiased estimator of (37) is given by

$$(40) \qquad \widehat{Y_{ie}Y_{if}} - \hat{Y}_{ie}\hat{Y}_{if}$$

where $\widehat{Y_{ie}Y_{if}} = cY_{ie} + dY_{ie}^2$, $\hat{Y}_{ie} = X_{ie}\hat{\beta}_{ie}$, and $\hat{Y}_{if} = X_{if}\hat{\beta}_{if}$.[12] The estimated Aitken matrix composed of terms such as (39) and (40) is then used in (38) to provide estimates of the multivariate multiple regression coefficients.

---

[11] $E\{\hat{Y}_{ie}(1 - \hat{Y}_{ie})\} = E\{(X_{ie}\beta_e + \varepsilon_e)(1 - X_{ie}\beta_e - \varepsilon_e)\}$
$= E\{X_{ie}\beta_e - (X_{ie}\beta_e)^2 - X_{ie}\beta_e\varepsilon_e + \varepsilon_e - X_{ie}\beta_e\varepsilon_e - (\varepsilon_e)^2\}$
$= E\{X_{ie}\beta_e(1 - X_{ie}\beta_e)\}$, assuming $E\varepsilon_e = 0$
$= E\{EY_{ie}(1 - EY_{ie})\}$
$= EY_{ie}(1 - EY_{ie})$.

[12] $Y_f = \gamma + \delta Y_e + \eta$.
Multiply each term in the true relation by the corresponding value of Y_e: $Y_eY_f = Y_e\gamma + Y_e^2\delta + \eta Y_e = (Y_e \mid Y_e^2)\begin{pmatrix}\gamma\\ \delta\end{pmatrix} + \eta Y_e$.
Then $\begin{pmatrix}c\\ d\end{pmatrix} = \begin{pmatrix}\gamma\\ \delta\end{pmatrix} + \left\{\left[\begin{pmatrix}Y_e'\\ Y_e^{2\prime}\end{pmatrix}(Y_e \mid Y_e^2)\right]^{-1}\begin{pmatrix}Y_e'\\ Y_e^{2\prime}\end{pmatrix}\right\}\eta Y_e$
$= \begin{pmatrix}\gamma\\ \delta\end{pmatrix} + H(Y_e)\,\eta Y_e$
$E\begin{pmatrix}c\\ d\end{pmatrix} = \begin{pmatrix}\gamma\\ \delta\end{pmatrix} + E\{H(Y_e)\eta Y_e\} = \begin{pmatrix}\gamma\\ \delta\end{pmatrix} + E_{Y_e}\{H\,H(Y_e)\eta Y_e \mid Y_e\}\}$
$= \begin{pmatrix}\gamma\\ \delta\end{pmatrix} + H(Y_e)\,E\{\eta Y_e \mid Y_e\} = \begin{pmatrix}\gamma\\ \delta\end{pmatrix}$.
Therefore, $EY_{ie}Y_{if} = E\{cY_{ie} + dY_{ie}^2\} = \gamma EY_{ie} + \delta EY_{ie}^2 = Y_{ie}Y_{if}$.
In non-linear cases the results are similar.

V. EMPIRICAL ILLUSTRATIONS

In this section the results of two empirical examples using microeconomic cross-section data are presented purely for illustrative purposes. Only a cursory attempt is made to provide any justification for the models postulated. The first example is meant to illustrate the case of a dichotomous dependent variable in which the proportions in the two categories are approximately equal. The second example also involves a dichotomous dependent variable in which the proportions are highly unequal

In both cases at least some regression coefficient and/or standard error was substantially altered by using GLS rather than OLS estimation. The R^2 rose in both cases, as the remark on page 9 suggested. The greatest effects were observed in the case where the proportions were highly unequal. To the best of my knowledge, the only empirical example in the literature is that of Zellner and Lee [13] who also found considerable variation in estimated standard errors.

Example a. Completion of a Training Course by Hard-Core Unemployed

The subjects for this example were 144 hard-core unemployed mean who were retrained at one of the "Big Three" automobile manufacturers in Detroit, Michigan, U.S.A. in 1969. The training course was approximately eight weeks in length. There were 79 completers and 65 dropouts for whom complete data were available. The actual completion rate, however, was substantially higher. Of a sample of 90 trainees, 79 completed the course and 11 did not. The other 54 cases in the sample represent dropouts from a previous training class who were interviewed at the same time as the other 90.

The question asked is: do those background factors of the individual trainee which employers traditionally use for selection purposes make a significant contribution toward explaining who are the completers and who the dropouts? The dependent variable is whether or not the man completed the training course. The code 1 was used for completers, 0 for dropouts. It was

---

[13] Arnold Zellner and Tong Hun Lee, "Joint Estimation of Relationships Involving Discrete Random Variables," Econometrica, April, 1965

hypothesized that the following independent variables should be of significance:

EDUC	Highest year of schooling completed according to company records
AGE	According to company records
PSTTNG	Dummy variable representing previous job training; coded 0 if respondent received no previous job training, 1 if he did
UN12MO	Number of months unemployed in the last 12 according to respondent's report

All respondents were male, black, and certified as hard-core by state officials. The results were:

Independent Variable	Estimated Coeff. OLS (Standard Error)	Estimated Coeff. GLS (Standard Error)
EDUC	-1.12 (2.30)	-1.54 (2.16)
AGE	1.06 (.60)	1.12 (.52)
PSTTNG	10.08 (9.44)	9.63 (9.46)
UN12MO	- .22 (1.17)	- .20 (1.15)
C (constant)	34.12 (30.23)	37.47 (29.49)
R^2	.0395	.1198

The results are expressed in terms of the number of percentage points effect an increase of one in the independent variable would have on the dependent variable.

Perhaps the most noticeable outcome is the total lack of statistical significance of the results. The R-squared increased by a factor of three. However, the t-statistics (ratio of estimated coefficient to standard error) were all less than two indicating no significant linear relation between the dependent variable and any of the explanatory variables at the 5% level. In percentage terms, some of the coefficients changed markedly. On balance, though, when the sample was about evenly divided between completers and dropouts, the heteroscedasticity correction had only a small effect on the results.

Example b. Successful On-the-Job Training of the Hard-Core

The subjects for this example were 73 hard-core unemployed men who were retrained in several large firms in Toledo, Ohio, U.S.A. in 1969. Successful retraining was defined according to length of time on the job and/or type of termination. A man was considered successfully retrained if he remained on the job for six months or more, left to take another job, or left to return to school. He was considered unsuccessfully retrained if he quit or was fired for job-related reasons. Three additional persons were omitted from the analysis. Successes were coded 1, non-successes 0. 87% of the trainees were so classified as successes.

The independent variables were:

EDUC Highest year of schooling completed

AGE

UN12MO Number of weeks unemployed in the last year

FAMINC Family income last year.

In terms of percentage point effect on the dependent variable, the results were:

Independent Variable	OLS	GLS
EDUC	.16 (.41)	.16 (.31)
AGE	.83 (.75)	.51 (.59)
UN12MO	.03 (.34)	.10 (.38)
FAMINC	-.004 (.003)	-.002 (.003)
C	72.33 (22.96)	73.25 (23.08)
R^2	.0495	.6682

As in the previous example, those background factors which employers traditionally look at made no statistically significant contribution toward explaining the variance in successful retraining. In contrast, the heteroscedasticity correction had considerable impact on a number of magnitudes. The coefficient on AGE was reduced by more than a third, the coefficient on UN12MO tripled, and the coefficient was cut in half. The standard errors on each variable were noticeably affected, generally smaller. But the model nevertheless failed to give any statistically significant results.

VI. CONCLUSION

This paper began with consideration of estimation procedures in a single-equation model in which the dependent variable is categorical. Whether there were two, three, or more possible values, the nature of the heteroscedasticity and the appropriate correction are identical. Although ordinary least squares (OLS) was shown to yield unbiased estimates of the regression coefficients, more efficient estimates could be obtained by Aitken (GLS) estimation which takes into account the heteroscedasticity.

The discussion was then extended to a model composed of a number of such equations. One possible estimation procedure is Aitken estimation equation-by-equation. However, Zellner's multivariate multiple regression estimation method is more efficient. Zellner and Lee's application of this method to systems of dichotomous dependent variables was then outlined and extended to systems of multichotomous dependent variables. This method was somewhat modified to handle situations in which the grouping of cases was impossible, so that an alternative procedure is suggested for estimating the intersection expectations needed for the Zellner-Lee estimator.

The Zellner-Lee procedure requires the inversion of a matrix which is GE x GE, where G is the number of groups and E the number of equations in the system. In the Zellner-Lee example of twelve groups and two equations, the required matrix is 24 x 24 which, although costly to invert, is still possible. However, consider for example a case of 500 observations for which grouping is impossible and three equations. The inversion of a 1500 x 1500 matrix not only exceeds the capacity of most (if not all) computers but would be prohibitively expensive. Inversion by partitioning, involving nine 500 x 500 diagonal matrices, would be possible but would require extensive manipulation of the resulting sub-matrices. Of course, the more equations, the more complicated the inversion procedure. The gain in efficiency of the Zellner-Lee procedure must be weighed against the costs in both human and computer time of achieving a more efficient estimate, considering that Aitken estimation equation-by-equation yields unbiased estimates which are in turn more efficient than OLS equation-by-equation.